Wolfgang Teubert / Cyril Belica

VON DER LINGUISTISCHEN DATENVERARBEITUNG AM IDS ZUR „MANNHEIMER SCHULE DER KORPUSLINGUISTIK"

Im ersten Teil („Linguistische Datenverarbeitung am Institut für Deutsche Sprache") berichtet Wolfgang Teubert, der diesen Bereich von 1975 bis 1993 leitete, von den ersten Arbeiten an und um Korpora herum – von den Anfängen in den sechziger Jahren des letzten Jahrhunderts bis in die neunziger Jahre, die einen Einschnitt in diese Historie auf mehrfache Art darstellen u.a. durch den Umzug des IDS an den heutigen Standort, durch neue Aufgabenverteilungen innerhalb des IDS und letztlich durch den Wechsel von Wolfgang Teubert an die Universität Birmingham im Jahr 2000. Im zweiten Teil („Die moderne Mannheimer Schule der Korpuslinguistik") konzentriert sich Cyril Belica, der lange Jahre zunächst der Arbeitsgruppe für Korpustechnologie und später dem Programmbereich Korpuslinguistik vorstand, auf die heutige korpuslinguistische Philosophie des IDS vor dem Hintergrund der Entwicklung der letzten zwanzig Jahre.

Linguistische Datenverarbeitung am Institut für Deutsche Sprache

Einer der Hauptgründe, warum es 1964 zur Gründung des Instituts für deutsche Sprache (so die damalige Schreibweise) kam, war der Versuch der Bewältigung einer Vergangenheit, in der sich weite Teile der germanistischen Sprachwissenschaft in ihrer Annäherung an nationalsozialistische Ideen vom positivistischen Vorbild einer wertneutralen, objektivistischen Wissenschaftlichkeit abgewendet hatten. Ein deklarativer Verzicht auf ideologische Anfechtungen und das Bestehen auf strikter Empirie schien auch deshalb geboten, weil es andernorts, in der Deutschen Demokratischen Republik, unübersehbare Bestrebungen gab, die Sprachgermanistik wie auch die anderen Geisteswissenschaften in den ideologischen Dienst des Staates zu stellen. In Westdeutschland glaubte man, die Verstrickung der Germanistik in alles Ideologische am besten dadurch zu vermeiden, dass man die Linguistik in den Rang einer ‚harten' Wissenschaft erhob, vergleichbar mit Physik oder Chemie. So wünschte sich der akademische Zeitgeist eine Rückkehr zu den Idealen einer Aufklärung, die bestrebt war, nicht nur natürliche sondern auch gesellschaftliche Vorgänge (einschließlich der gesellschaftlichen Praxis von Kommunikation durch Sprache) als Systeme autonomer, kausal definierter Gesetzmäßigkeiten zu begreifen, ohne mehr als Lippenbekenntnisse zu dem anderen Aspekt der Aufklärung, nämlich der Emanzipation des Menschen als einer autonomen Person, abzulegen. Denn auch eine sich lediglich deskriptiv verstehende Sprachwissenschaft, wie sie damals unter Berufung

auf den Strukturalismus von Ferdinand de Saussure als ideologisch unverdächtig importiert wurde, verneint ja die kreative Freiheit der Sprecher und zwängt sie in ein Korsett eines synchronisch verstandenen Sprachsystems, dessen Gesetze nicht weniger permanent (d.h. änderungsresistent) sind als die Naturgesetze, auf die also die Kommunikationsteilnehmer keinen Einfluss haben.

Doch um die Regeln zu erkunden, die einem so verstandenen Sprachsystem unterliegen, bedurfte es erst einmal einer Bestandsaufnahme. Während auf der anderen Seite des Atlantiks Noam Chomsky Sprache auf ein rein biologisches Phänomen reduzierte, in dem ein angeborenes Sprachorgan für die Generierung grammatikalischer Sätze sorgte, so dass von den Äußerungen der Sprecher nur das als systemkonform gilt, was diesen hypostasierten Gesetzmäßigkeiten genügt, schien wenigstens den angewandten Sprachgermanisten eine solche Beschäftigung mit dergleichen Sprachuniversalien als unergiebig, waren sie doch im kleinräumigen Europa mit einer Sprachenvielfalt konfrontiert, die es angeraten zu lassen schien, den Schwerpunkt der Arbeit eher auf solche Aspekte zu legen, die Sprachen voneinander unterscheiden. Für sie waren Sprachen nicht so sehr biologische als vielmehr soziale Phänomene, denen man nicht durch Introspektion, sondern nur durch Datenerhebung auf den Grund gehen konnte.

Einen zusätzlichen Anstoß erhielten solche neopositivistischen Ideen von der plötzlichen Ubiquität der Elektronengehirne. Seit den fünfziger Jahren des letzten Jahrhunderts wurde fest daran geglaubt, dass nun bald all das, was dem menschlichen Geist immer noch widersprüchlich, chaotisch und zufällig erschien, in einer systematischen Ordnung aufgehen würde, die uns nur die unbegrenzten Kombinationsmöglichkeiten dieser Wunderwerke vor Augen zu führen vermögen. Also galt es, Daten zu sammeln. Inwieweit es bei der Gründung des IDS den im Kuratorium vertretenen Ordinarien bewusst war, dass im Jahr zuvor, an der amerikanischen Brown University, das mit einer Million laufender Textwörter erste größere maschinenlesbare Korpus der Welt, das sogenannte Brown Corpus, aufgebaut von W. Nelson Francis und Henry Kucera, das Licht der Welt erblickt hatte, ist nicht bekannt. Aber dass der Gedanke, sich für sprachwissenschaftliche Fragestellungen des Computers zu bedienen, damals in der Luft lag, zeigt sich auch darin, dass zur selben Zeit in Edinburgh auch John Sinclair, der Nestor der Korpuslinguistik, sein erstes Korpus schuf und es sogleich dazu benutzte, mit liebgewordenen Axiomen der Sprachwissenschaftler aufzuräumen, etwa dass das Einzelwort in Isolierung den zentralen Platz in der Bedeutungslehre einnimmt. Auch wenn das Prinzip der Kollokation bereits dreißig Jahre früher von A.S. Hornby und Harold Palmer erkannt worden war, bedurfte es des Computers, um zu zeigen, wie stark davon sprachliche Sinneinheiten geprägt sind.

Eine solche Revolutionierung der Linguistik lag den Kuratoren indessen fern. Ihren Wünschen entsprechend sollten die wissenschaftlichen Mitarbeiter des IDS lieber an die Tradition der deutschen Sprachwissenschaft, mit den Brüdern Grimm, August Schleicher, Hermann Paul, Otto Behagel, anknüpfen, die auch ohne Computer bereits datenbezogen an einer systematischen Analyse der deutschen Sprache gearbeitet hatten, und ihre Modernität durch die Verwendung der nun verfügbaren Technologie unter Beweis stellen. So heißt es im ersten Jahrbuch des Instituts für deutsche Sprache (IDS (Hg.) 1967, S. 12):

Die Abteilung Dokumentation des heutigen Deutsch' soll ein vordringliches Bedürfnis erfüllen. In einem repräsentativen Querschnitt durch das heutige Schrifttum und aufgrund einer sorgsam vorbereiteten Programmierung soll zunächst die geschriebene deutsche Sprache von heute auf elektronischem Wege gespeichert werden, wobei Gesichtspunkte der Wortlehre und des Satzbaus berücksichtigt werden. Das so gewonnene Material soll allen Forschern, die sich um die deutsche Sprache bemühen, zur Verfügung stehen.

Allerdings waren Elektronengehirne damals noch dünn gesät. Das nächstgelegene war im Rechenzentrum Darmstadt zu finden. Nun galt es, ein, wie man damals meinte, repräsentatives Korpus der deutschen Gegenwartssprache zusammenzustellen und es dann auf ausgesprochen arbeitsintensive Weise über Lochstreifen zu erfassen. Diese wurden dann nach Darmstadt gebracht und eingelesen und lagen sodann maschinenlesbar vor. Die ausgedruckten Daten wurden anschließend wieder in Mannheim korrigiert.
Die erste elektronische Textsammlung des IDS war das Mannheimer Korpus (MK1), bestehend aus 293 Texten oder 2,2 Millionen Wörtern, die zwischen den frühen fünfziger Jahren und 1967 veröffentlicht worden waren, mit dem Schwerpunkt auf Literatur, aber auch mit einigen populärwissenschaftlichen Büchern sowie Texten aus Zeitungen und Illustrierten. Die Debatte um Repräsentativität oder Ausgewogenheit von Korpora hatte gerade erst begonnen. Aber schon bald wurden Defizite empfunden, die durch das sehr viel kleinere Mannheimer Korpus 2 ausgeglichen werden sollten, das bisher übersehene Textsorten wie Broschüren und Heftchenromane enthielt. Ende der sechziger Jahre war also das Mannheimer Korpus verfügbar. Allerdings wurde es anfänglich nur recht wenig genutzt. Erklärlich war das, weil jeder Versuch, bestimmte Daten zu extrahieren, die ein grammatisches Phänomen hätten beleuchten können, ausgesprochen umständlich war. Erst musste das Problem formal beschrieben werden. Das bedeutete mühevolle Kleinarbeit. Wollte man etwa wissen, welche Verben an die Stelle von Genitivobjekten auch Präpositionalobjekte zulassen („er erinnerte sich ihrer" neben „er erinnerte sich an sie"), musste eine Liste nicht nur aller möglichen Verben son-

dern auch aller ihrer Wortformen erstellt werden. Ein Computerprogramm musste geschrieben und auf Lochkarten eingestanzt werden, es musste nach Darmstadt gebracht und dort von den Spezialisten zum Laufen gebracht werden (was meist erst nach vielen Korrekturen der Fall war). Dann konnte der Rechner einen Ausdruck liefern, der eine Konkordanz aller gefundenen Belege enthielt. Diese musste dann manuell ausgedeutet werden. Denn ob sich in einem entsprechenden Satz nun ein Genitivobjekt oder ein Präpositionalobjekt findet, können Computer selbst heute, wo sie mit umfangreichen und komplexen Hintergrundprogrammen arbeiten, nur annäherungsweise erkennen. Ein weiteres Problem ist die Ambiguität von Wortformen. In dem Satz „Er erklärte, dass sie das Ziel erreicht haben." kann beispielsweise *haben* sowohl als Indikativ als auch als Konjunktiv interpretiert werden.

Es ist dies ein heute immer noch gern verdrängtes Grundproblem in weiten Teilen der Korpuslinguistik, dass sie ihrem Anspruch, in ihren Verfahren auf einer Ebene mit den (idealisierten) Naturwissenschaften zu stehen, nicht gerecht werden kann. Zwar kann sie nachprüfbare Ergebnisse bereitstellen, aber diese Ergebnisse bedürfen der (manuellen – vielleicht sollte man eher sagen: der intellektuellen) Interpretation durch den Wissenschaftler. Auch heute noch wird darüber diskutiert, ob das korpuslinguistische Verfahren eher korpusbasiert oder korpusgesteuert sein sollte. In ersterem Fall benutzt die Analyse vorgegebene Kategorien (etwa, was ein Objekt ist) und die manuelle Annotation der Ergebnisdaten. Im zweiten Fall sollen sich die relevanten Kategorien allein aus dem Kookkurrenzverhalten der Daten (im wesentlichen Zeichenketten zwischen Leerzeichen) ergeben. Bestimmte Wörter oder Wortformen kommen in statistisch signifikanter Weise immer wieder miteinander vor. Beispielsweise *ertränkt* man seinen *Kummer* über eine Nachricht gern in Alkohol, während aus *Gram* über vielleicht dieselbe Nachricht auffällig häufig *gestorben* wird. Doch kann es mit dieser Erkenntnis wohl kaum sein Bewenden haben. Doch wie sich die Bedeutung von *Gram* und *Kummer* genau unterscheidet, kann uns der Computer nicht sagen; die Analyseergebnisse bedürfen vielmehr der Interpretation. Dafür indessen gibt es keine „wissenschaftliche" Methode.

Das Projekt *Grundstrukturen der deutschen Sprache* war das erste große Vorhaben, in dem das Mannheimer Korpus ausgewertet wurde. Von 1971 bis 1981 erschienen siebzehn Bände, die sich mit Themen wie dem *Konjunktiv in der deutschen Sprache der Gegenwart* (Siegfried Jäger), *Satzbauplan und Wortfeld* (Bernhard Engelen), *Wortstellung* (Ursula Hoberg) oder dem *Passiv im heutigen Deutsch* (Klaus Brinker) beschäftigten. Nicht nur wegen ihrer Korpusbezogenheit, sondern auch und gerade wegen ihrer interpretativen Schlussfolgerungen waren diese Arbeiten erfreulich innovativ. Andere korpusbasierte Arbeiten, die Neuland erschlossen, erschienen in der Reihe *Sprache der Gegenwart*. Es waren ausgesprochen produktive Jahre, nicht nur in Hinblick auf die erziel-

ten Ergebnisse, sondern auch im Sinne der Theoriebildung. Dazu gingen die Meinungen weit auseinander. Es gab keinen Methodenzwang und ebenso wenig eine IDS-Sprachphilosophie. Die damalige Institutsleitung, besonders Ulrich Engel, erst stellvertretender Direktor (neben Paul Grebe) und später einziger Direktor, sah sich mit einem durchaus fruchtbaren Klima akademischer Auseinandersetzung konfrontiert, in dem alle möglichen Ansätze durchaus kontrovers diskutiert wurden, ohne dass sich die Mitarbeiter auf ein bestimmtes Modell festlegen ließen.

Im Jahr 1969 schaffte sich das IDS endlich seine eigene Rechenanlage an, die, wie es damals der Fall war, ein ganzes Geschoss belegte, aber in Hinblick auf Datenzugriff und -manipulation keine große Erleichterung brachte. Ein Rechenzentrum wurde eingerichtet, dessen Mitarbeiter zwar die Wünsche der Forschungsabteilungen erfüllen sollten, die sich aber lieber, wie damals in Rechenzentren allgemein üblich, ganz vorwiegend ihren eigenen Prioritäten widmeten. Bis in die achtziger Jahre musste umständlich *offline* mit Ausdrucken gearbeitet werden, die von den wissenschaftlichen Mitarbeitern manuell ausgewertet wurden. Im Wesentlichen handelte es sich dabei um Konkordanzen von Listen von Wortformen. Selbst die heute so unentbehrliche Anwendung von Statistikprogrammen zur Untersuchung von Kookkurrenzen steckte damals noch in den Kinderschuhen. Das Korpus wurde oft eher als Steinbruch für Belegstellen und nicht im Sinne einer systematischen Datenanalyse genutzt. Andere Korpuszentren, etwa in Birmingham, in Oslo oder in Göteborg, betrieben damals schon mehr von dem, was man heute unter Korpuslinguistik versteht. Ob es der Forschung am IDS gut getan hätte, wenn die selbstreferenzielle Autokratie des Rechenzentrums aufgebrochen worden wäre, mag dahingestellt bleiben. Denn der Kanon der Themen, die sich korpuslinguistisch gewinnbringend bearbeiten ließen, war seinerzeit doch recht begrenzt. So hat die an sich unbefriedigende Rechenzentrumssituation vielleicht sogar der Forschungsvielfalt des IDS gut getan.

Inzwischen war aus der Außenstelle des IDS in Freiburg ein Korpus gesprochener Sprache nach Mannheim gelangt (das Freiburger Korpus), das als Datengrundlage für Monographien zu Phänomenen gesprochener Sprache herangezogen wurde. Auch damit hat das IDS Neuland betreten. Zu erwähnen ist auch das den damaligen politischen Verhältnissen geschuldete Projekt des *Bonner Zeitungskorpus*, das von Manfred Hellmann in einer Bonner Außenstelle des IDS aus den Zeitungen *Die Welt* und *Neues Deutschland* kompiliert wurde, um den Nachweis zu erbringen, dass die DDR dabei war, die sprachliche Einheit Deutschlands zu zerstören. Eine auf diesem Korpus fußende lexikografische Auswertung wurde allerdings erst Ende der neunziger Jahre abgeschlossen, als die Spaltung Deutschlands längst Geschichte geworden war.

Es soll hier nicht verschwiegen werden, dass die linguistische Datenverarbeitung am IDS zuweilen Ausflüge in entlegenere Gefilde der Verarbeitung natürlicher Sprache unternommen hat. In der zweiten Hälfte der siebziger Jahre hat kaum eine Erwartung die Menschen mehr in Atem gehalten als das Konzept der Künstlichen Intelligenz, dessen Realisierung man sich von der jeweils nächsten Computergeneration versprach. Die Drohung einer japanischen Überlegenheit auf diesem Gebiet (die sich als völlig unzutreffend herausstellen würde) ließ nicht nur in Deutschland die Mittel dafür reichlich fließen. Das IDS wurde mit dem Projekt PLIDIS (‚Problemlösendes Informationssystem mit Deutsch als Interaktionssprache') in diese Anstrengungen eingebunden. Geleitet wurde es von Genevieve Berry-Rogghe und Gisela Zifonun. Speziell ging es um ein Verfahren, das es menschlichen Nutzern gestatten würde, in natürlicher Sprache, also auf Deutsch, mit einer landesweiten Datenbank mit Informationen zu Industrieabwässern zu interagieren. In diesem Zusammenhang wollten die Leiterinnen dieses Projekts auch eine formalistische Grammatiktheorie (die Montague-Grammatik) testen, die sprachliche Äußerungen in formallogische Ausdrücke übersetzen würde, die von einem entsprechenden Computerprogramm verarbeitet werden könnten. Wie eigentlich alle dieser Projekte weltweit ist auch dieses Vorhaben einerseits daran gescheitert, dass es sich bei ‚natürlicher' Sprache um eine kontingente kulturelle Errungenschaft handelt, die sich nicht in das Korsett formallogischer Regeln fassen lässt, andererseits daran, dass beim fördernden Ministerium schließlich die Geldknappheit die Furcht vor japanischer Überlegenheit überwog. Externer Leiter des Bereichs Linguistische Datenverarbeitung war zu dieser Zeit der früh verstorbene Kommunikationswissenschaftler Gerold Ungeheuer, dessen Thema damals die Kommunikation zwischen Mensch und Maschine war. Auch in dem Projekt PLIDIS zeigt sich die damals noch weithin ungebändigte Kreativität des Instituts, dessen Mitarbeiter sich noch mehr als Denkfabrik denn als Dienstleister für die internationale Sprachgermanistik verstanden haben.

Die Arbeitsstelle Linguistische Datenverarbeitung (LDV) widmete sich in ihrer nächsten Phase wieder ganz überwiegend der Bereitstellung von Korpora und der Verbesserung von Zugriffsmöglichkeiten auf diese Daten. So wurde Anfang der achtziger Jahre etwa der Wortformengenerator MOLEX verwirklicht, der zu über 60.000 Wörtern alle Flexionsformen generiert und ihre morphosyntaktische Funktion ausweist, womit sich für die Korpusnutzer der Datenzugriff ganz erheblich erleichterte.

Doch mit seinem pluralen Enthusiasmus für theoretische und methodische Innovation geriet das IDS immer mehr in Konkurrenz zur Sprachgermanistik an den Universitäten. Frei von Verwaltungsaufgaben und Lehrverpflichtungen sonnten sich die IDS-Mitarbeiter in gleichsam paradiesischen Zustän-

den, ganz überwiegend dauerhaft vertraglich abgesichert durch gesicherte staatliche Zuwendungen, von denen der schon damals prekäre universitäre Mittelbau nur träumen konnte. Doch so unabhängig, wie sich die IDS-Wissenschaftler wähnten, waren sie in Wirklichkeit nicht. Über ihnen thronte das Kuratorium, in dem außer den geldgebenden Staatsvertretern führende professorale Vertreter der germanistischen Sprachwissenschaft über die Geschicke des IDS befanden sowie seine Ziele und Aufgaben diktierten. Als das durch drittmittelfinanzierte Projekte stark angewachsene Institut immer mehr als mögliche Bedrohung der professoralen Meinungsführerschaft angesehen wurde, nahm man eine durch die Unberechenbarkeit öffentlicher Projektförderung verursachte finanzielle Krise zum Anlass, die Leitung auszuwechseln und zum Nutzen der in- und ausländischen Sprachgermanistik die Richtung weg von explorativen Projekten hin zu der Bereitstellung von Ressourcen und zu Vorhaben zu lenken, in denen eine größere Zahl von Mitarbeitern über mehrere Jahre hinweg umfangreiche Datenbestände erfassen und aufarbeiten würden. Das galt ganz besonders auch für die linguistische Datenverarbeitung am Institut. Die Beschaffung von Daten, also die Kompilation neuer Korpora besonders in Hinblick auf die unmittelbare Gegenwartssprache, für die drei Hauptarbeitsbereiche Lexik, Grammatik und Pragmatik sowie die optimale Gestaltung des Datenzugriffs standen von nun an im Mittelpunkt der Arbeit. Andererseits sollte keine Sprachakademie entstehen. Was dem Kuratorium vorschwebte, war, das IDS in ein Dienstleistungszentrum für die germanistische Sprachwissenschaft im In- und Ausland umzuwandeln.

Für den Bereich Linguistische Datenverarbeitung hätte die Entwicklung in der Tat ein Schritt in die richtige Richtung sein können. Leider kam es durch problematische Entscheidungen damals noch nicht zu dem erhofften Durchbruch der Korpuslinguistik am IDS. Zu einer Zeit, als die Welt längst auf weitestgehend dezentralisierte Datenverarbeitung umgestellt hatte, mussten die Forscher am IDS immer noch um ihren Rechnerzugriff kämpfen. Denn Rechenzentren waren immer und überall auf Zentralisierung erpicht, weshalb sie eine *mainframe*-Konfiguration bevorzugten, die den Nutzern allenfalls Bildschirmterminals zustanden, nicht aber ihre eigenen PCs. Als Folge verzichteten viele Projekte lange auf die Entwicklung und den Einsatz echter korpuslinguistischer Verfahren. Erst neu hinzu gekommenen Mitarbeitern war es zu verdanken, dass nicht nur die Korpusbasis enorm ausgeweitet wurde, sondern dass nun auch endlich ein versatiles und flexibles Zugriffssystem (COSMAS, s.u.) entwickelt wurde, das wirklich an der Spitze internationaler Korpustechnologie stand. Der COSMAS-Nutzer musste nun seine Wünsche nicht mehr mit Ansprechpartnern in der LDV aushandeln, sondern konnte seine Korpusanfragen, komplex wie sie sein mochten, selbst konfigurieren.

Damit wurde endlich korpusbasiertes Arbeiten in den wesentlichen Projekten im Bereich Lexik zur gern genutzten Gewohnheit. Ein ganz besonderer Pluspunkt von COSMAS war und ist, dass es nicht nur Mitarbeitern, sondern auch den Gästen des Instituts und auch über das Internet interessierten Korpuslinguisten überall in der Welt zur Verfügung stand und immer noch steht. Wenngleich nun in vielen Projekten, vor allem im Bereich Lexik, die IDS-Korpora die wesentliche Datenquelle darstellten, hatten die Nutzer nur wenig Kenntnis von den Diskussionen in der schnell anwachsenden globalen Gemeinde von Korpuslinguisten. Denn diese Gruppe von Sprachforschern kommunizierte auf Englisch, und Germanisten hatten lange Bedenken, ihre eigene Sprache, das Deutsche, zu marginalisieren, indem sie über ihre Forschungen auf Englisch verhandelten. Anfänglich war es auch durchaus so, dass sich Korpuslinguisten im Ausland vor allem mit dem Englischen beschäftigten. Sie hatten schon 1979 begonnen, sich zu organisieren, in einer Organisation namens ICAME (International Computer Archive of Modern and Medieval English). Schon im Titel wird sichtbar, dass sich dieser Ansatz, wo auch immer er betrieben wurde, fest in den Händen der Anglistik befand, und das ist großenteils noch heute der Fall. Inzwischen wird die Beschäftigung mit anderen Sprachen indessen durchaus begrüßt. Doch noch immer versteht sich auch die internationale Korpuslinguistik ganz überwiegend als Teil der *applied linguistics*, wie sie sich in Großbritannien und den nach britischem Vorbild organisierten Anglistiken in anderen Ländern herausgebildet hat, also zuvorderst als Hilfsdisziplin für den Fremdsprachenunterricht und die Lexikografie (vor allem Lernerwörterbücher). Es ist daher nicht weiter verwunderlich, dass auch die IDS-Mitarbeiter, die mit Korpora arbeiteten, sich nicht in dieser Gemeinde verortet sehen wollten.

Erst in der Mitte der achtziger Jahre begann sich die linguistische Datenverarbeitung am IDS vor allem in ihren korpusbezogenen Aspekten in das europäische Geflecht vergleichbarer Aktivitäten zu integrieren. Die Anregungen dazu kamen von außen. Denn inzwischen hatte auch die Computerlinguistik, der es um technische Anwendungen im Bereich automatische Sprachverarbeitung geht, den Nutzen von Korpora erkannt. Sie versprach den europäischen Politikern, Lösungen für die multilinguale Problematik Europas zu liefern. Die Europäische Union suchte zu dieser Zeit ihre Identität in der Verbindung einer einheitlichen, auf denselben Grundrechten basierenden demokratisch organisierten Zivilgesellschaft, die zugleich ihre kulturelle und sprachliche Vielfalt förderte. Aber damit eine einheitliche Zivilgesellschaft entstehen kann, müssen ihre Mitglieder miteinander kommunizieren können. Damit ein vielsprachiges Europa funktionieren konnte, galt es also, die Voraussetzungen für eine Kommunikation über Sprachgrenzen hinweg zu schaffen. Zunächst war es der Europarat, der kleinere Kooperationsprojekte

unterstützte. So nahm das IDS an einem multilingualen korpusbasierten Vorhaben teil, das von John Sinclair geleitet wurde und an dem Korpuslinguisten aus Jugoslawien, Italien, Schweden und Spanien beteiligt waren. Dabei schälte sich als wünschenswertes Ziel die Schaffung eines Netzwerks europäischer Referenzkorpora heraus, als ausreichend große Korpora (mit jeweils über hundert Millionen Wörtern), die einheitlich zu kodieren waren, damit lexikalische und grammatische Phänomene in den befassten Sprachen miteinander verglichen werden konnten. Dieses Projekt (NERC: Network of European Reference Corpora) wurde bereits von der Europäischen Kommission finanziert. Das IDS war darin für Deutschland vertreten. Es legte eine *roadmap* (wie man heute sagen würde) für die weitere Zusammenarbeit fest, und es definierte auch die Parameter für Korpuskompilation und Korpuskodierung. Das Anschlussprojekt PAROLE (‚Preparatory Action for Linguistic Resources Organization for Language Engineering'), das eigentlich der Schaffung dieser Referenzkorpora dienen sollte, war weniger erfolgreich, was auch an den unbefriedigenden Rahmenbedingungen lag.

Ende der achtziger Jahre kam es dann bekanntlich zum Zusammenbruch der politischen Systeme in Mittel- und Osteuropa. Dort hatten sich über viele Jahre hinweg an Universitäten und mehr noch an Akademien wichtige Zentren herausgebildet, in denen linguistische Datenverarbeitung auf hohem Niveau betrieben wurde. Um die dort tätigen Wissenschaftler in die (west)europäische Forschung einzubinden, finanzierte Brüssel unter dem Stichwort *human language technology* (HLT) zahlreiche paneuropäische Projekte, die teilweise der Schaffung einer Infrastruktur, teilweise der Schaffung multilingualer Ressourcen (vor allem Korpora und Korpussoftware) dienten. In vielen dieser Vorhaben war das IDS die deutsche Partnerinstitution, so beispielsweise in PAROLE II, ELAN, SIMPLE und MECOLB. Am IDS war auch die Projektleitung für das langlebige Projekt (1995-2002) TELRI (‚Trans-European Language Resources Infrastructure') mit Partnerinstituten in über zwanzig Ländern angesiedelt. Während solche Projekte letzten Endes wenige Ressourcen von bleibendem Wert schufen, zeigten sie doch die Bedeutung echter Sprachdaten für die Entwicklung von vermarktbarer Kommunikationstechnologie in einem multilingualen Rahmen. Darüber hinaus wiesen sie den Weg hin zu mehr Zusammenarbeit in der europäischen Sprachforschung. So leistete auch die linguistische Datenverarbeitung einen Beitrag dazu, dass sich das IDS in der Zusammenarbeit europäischer Sprachzentren profilierte. Auch öffnete sich zu dieser Zeit die traditionell britisch ausgerichtete Korpuslinguistik nicht nur vermehrt anderen Sprachen, sondern auch neuen Themen, die über Wörterbucharbeit und Fremdsprachenunterricht hinausreichten. Deshalb nahmen seit den neunziger Jahren wissenschaftliche Mitarbeiter zunehmend an internationalen Tagungen zum Thema Korpus-

linguistik teil und machten auf die Arbeit am IDS in englischsprachigen Beiträgen aufmerksam. Für die IDS-Forschung wichtiger war vielleicht, dass das korpuslinguistische Arbeiten die Barriere zwischen den Bereichen LDV und den Forschungsabteilungen beseitigte. Dazu trug auch die schrittweise Reorganisation des Instituts bei. Heute arbeitet eine neue Generation von Wissenschaftlern, seien sie nun von ihrer Ausbildung her Sprachwissenschaftler oder Informatiker, gemeinsam an neuen Ideen und konzipiert dazu innovative Verfahren, die weltweit als führend gelten können. Wurde von der linguistischen Datenverarbeitung am Institut am Anfang erwartet, den Anspruch der germanistischen Sprachwissenschaft auf Wissenschaftlichkeit im positivistischen Sinn zu untermauern, bekommt sie nun immer mehr eine neue Aufgabe, nämlich die technischen Voraussetzungen für die Analyse ganzer Diskurse zu schaffen. In einer Zeit, in der die Welt, in der wir uns befinden, immer seltener durch eigene Anschauung und immer mehr durch Sprache vermittelt wird, hilft die Korpuslinguistik, die sprachlichen Konstrukte durchsichtig zu machen, in denen sich unterschiedliche Interessen manifestieren. In diesem Sinn kann die Sprachgermanistik heute einen Beitrag zum Selbstverständnis der Gesellschaft leisten, der sie angehört. Das Institut für Deutsche Sprache, das sich jetzt zunehmend auch dieser Aufgabe stellt, hat sich in den fünfzig Jahren seines Bestehens unentbehrlich gemacht.

Die moderne Mannheimer Schule der Korpuslinguistik

DeReKo – Das Deutsche Referenzkorpus

Der Umzug des IDS im Jahr 1992 in die Mannheimer Innenstadt markierte auch einen qualitativen Sprung für die Korpusinfrastruktur und -forschung. Da der alte zentrale Rechner Siemens BS2000 nicht in das neue Rechenzentrum in R 5 migriert werden konnte, war eine grundsätzliche Entscheidung

Abb. 1: Hardware für die Korpora 1993 und 2013, Zunahme der Speicherkapazität von 100 MB zu 10 TB um das ca. 100.000-fache

über die künftige technische Infrastruktur erforderlich und die – wie sich heute zeigt richtige – Wahl fiel auf Unix-basierte Systeme.
Da aber die Softwareentwicklung in der damaligen Arbeitsstelle Linguistische Datenverarbeitung, von der Datenerfassung bis hin zur Recherchesoftware REFER in den achtziger Jahren, eng an die BS2000-Architektur gebunden war, hatte dies zur Folge, dass der gesamte Bereich der Korpusarbeit am IDS neu konzipiert und das neue Konzept noch vor dem Umzug zumindest teilweise umgesetzt werden musste. Unter dem Namen COSMAS (*Corpus Search, Management and Analysis System*) wurde ein Projekt ins Leben gerufen, dessen Ansätze zur Methodik der Einbettung der Korpusbefragung in die linguistische Forschung in den nächsten Jahren die Vision einer in der Empirie fest verankerten Linguistik nicht nur in technischer, sondern vor allem in wissenschaftsmethodischer Hinsicht nachhaltig geprägt haben. Vor dem Hintergrund der neu zu entwickelnden Recherchesoftware wurde in Form einer integrativen Korpusplattform eine Reihe von innovativen korpuslinguistischen Prinzipien formuliert, die den gesamten Bogen von Korpusakquisitions- und -samplingstrategien, über das Konzept von virtuellen Korpora, des einheitlichen Datenmodells und der urheberrechtlichen Unbedenklichkeit, das Sinclairsche Prinzip der *minimum assumption* und der mehrfachen, konkurrierenden Schichten von linguistischen Annotationen, bis hin zur Korpusanalyse- und -erschließungsmethodik inklusive nachgelagerter (d.h. beim Ordnen der Evidenzen möglichst weit nach hinten verschobener) Interpretation umspannte. Schritt für Schritt wurden die vorhandenen IDS-Korpora in die neue Plattform integriert und es wurde mit der kontinuierlichen Akquisition von Texten für ein neuartiges, universelles Archiv des geschriebenen Deutsch begonnen, welches seit dem Abschluss des im Jahr 1998 initiierten drittmittelfinanzierten Kooperationsprojekts „DeReKo – Das Deutsche Referenzkorpus“ dessen Namen weiterführt.
Heute gehört DeReKo (IDS 2013) zu den weltweit größten linguistisch motivierten Archiven des geschriebenen Deutsch der Gegenwart. Es enthält 6 Milliarden (Stand: Juni 2013) Textwörter mit einer durchschnittlichen Wachstumsrate von ca. 300 Millionen Textwörtern pro Jahr.
Es ist als eine vielseitige empirische Basis für die sprachwissenschaftliche Forschung konzipiert, mit einer breiten stichprobenartigen Abdeckung des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland und, gleichwohl in weit geringerem Umfang, auch in Österreich und in der Schweiz.
An dieser Stelle sei allen ausdrücklich gedankt, die auf vielfältige Weise zum Aufbau des Archivs und der umgebenden Infrastruktur beigetragen haben (vgl. http://www.ids-mannheim.de/kl/projekte/korpora/ehemalige.html).

Zugang zu DeReKo

Aus lizenzrechtlichen Gründen können die meisten DeReKo-Texte nicht frei kopiert oder heruntergeladen werden, sondern stehen primär über das Recherche- und Analysesystem COSMAS II zur Verfügung. In COSMAS II sind derzeit mehr als 30.000 Nutzer aus über 100 Ländern registriert. Eine weitere Möglichkeit, DeReKo im Sinne der bestehenden Lizenzbedingungen zu nutzen, sieht vor, dass externe Wissenschaftler ihre Programme zur Analyse und Auswertung von DeReKo auf den Rechnern des IDS laufen lassen können, im Sinne der Maxime „wenn die Daten nicht zum Code dürfen, muss der Code zu den Daten kommen". In Projekten wie CLARIN (Váradi et al. 2008), D-SPIN (Bankhardt 2009) oder TextGrid (Gietz et al. 2006) werden außerdem Forschungsinfrastrukturen entwickelt, die es ermöglichen, derartige Dienstleistungen auch in Form von Web/Grid-Anwendungen anzubieten.

Korpusdesign und -komposition: Repräsentativität und Stratifikation

Der eigentliche Sinn einer Korpusbefragung liegt in unserer Erwartung, dass die im Korpus gemachten Beobachtungen sich auch auf einen Sprachausschnitt über das Korpus hinaus verallgemeinern lassen. Diese Erwartung ist aber nur dann berechtigt, wenn das Korpus eine ausreichend repräsentative Stichprobe des Sprachausschnitts darstellt, auf den unsere Beobachtungen extrapoliert werden sollen. Dies ist jedoch eine im Allgemeinen nur schwerlich zu erfüllende Bedingung, die nur in Ausnahmefällen hinreichend rigoros formuliert werden kann. Einer der methodischen Fallstricke besteht darin, dass genau dasjenige Objekt, das *vor* der Wahl eines geeigneten, repräsentativen Korpus definiert sein müsste, der gemeinte Sprachausschnitt, zugleich das „unbekannte" Objekt ist, welches *durch* die Korpusbefragung näher inspiziert und erschlossen werden soll. Daher wird oft eine iterative Vorgehensweise gewählt: Von einer ersten, annähernden Definition der Repräsentativität ausgehend versucht man die Plausibilität der Korpusbefunde einzuschätzen und, wenn nötig, die Repräsentativität des Korpus durch Anpassungen der Korpuszusammensetzung zu verbessern. Mit Hilfe der sog. *stratifizierten Stichprobenziehung* können grobe Fehler der Korpuszusammensetzung im Hinblick auf einen konkreten Sprachausschnitt vermieden werden, indem man den Sprachausschnitt in Gruppen – *Strata* – unterteilt, die für seine Eingrenzung besonders relevant sind, beispielsweise Modus, Genre, Texttyp, Thema, Publikum, Zeit oder Register. Man zieht dann Stichproben aus den einzelnen Strata, und zwar in einem vorher gewählten quantitativen Verhältnis zueinander.

Ausgewogene Stichproben und Urstichproben

Im Unterschied zu den meisten großen Korpora ist das DeReKo-Archiv nicht als eine – wie auch immer definierte – repräsentative Stichprobe des Sprachgebrauchs, sondern als ein universelles Textarchiv konzipiert, aus dem möglichst viele unterschiedliche, für verschiedene Problemstellungen jeweils repräsentative Stichproben, *virtuelle Korpora* genannt, schnell und bequem gezogen werden können. Das DeReKo-Archiv selbst versteht man dabei als eine Ur-Stichprobe (*primordial sample*, Kupietz et al. 2010) des allgemeinen Sprachgebrauchs, die vor allem im Hinblick auf ihre Größe und auf eine durch die stratifizierte Stichprobenziehung angestrebte größtmögliche Vielfalt des zur Verfügung gestellten Sprachmaterials kontinuierlich ausgebaut wird. Diese Herangehensweise, nämlich maßgeschneiderte, problemspezifische virtuelle Korpora dynamisch aus einer vorliegenden Ur-Stichprobe bilden und bei Bedarf leicht optimieren zu können, hat gegenüber dem traditionellen Ansatz von statischen Korpora sowohl unter methodischen als auch unter wirtschaftlichen Gesichtspunkten viele Vorteile. Insbesondere die Flexibilität, mit der man sich mit Hilfe von DeReKo einer optimalen, für eine konkrete Fragestellung repräsentativen Korpuszusammensetzung nähern kann, und der Grad der Wiederverwendbarkeit des einmal akquirierten Textmaterials in verschiedenen Forschungszusammenhängen sind der Ansporn auch bei den aktuellen nationalen und internationalen Forschungsarbeiten zu *virtuellen Kollektionen* (van Uytvanck 2010), einer Verallgemeinerung von virtuellen Korpora und des Ur-Stichprobenprinzips auf beliebige Sprachressourcen.

Korpusgröße

Für die Generalisierbarkeit von Korpusauswertungen sowie überhaupt für erfolgreiche Korpusrecherchen ist eine weitere wichtige – wenn nicht sogar die wichtigste – Eigenschaft eines Korpus schlichtweg seine Größe. Es lässt sich nicht im Allgemeinen sagen, wie groß ein Korpus sein sollte (vgl. aber Biber 2008 für eine eingehendere Besprechung dieses Aspekts). Aber selbst im Zeitalter sehr großer Korpora bringt es nach wie vor Robert Mercers Forderung „more data is better data" (Church/Mercer 1993) auf den Punkt. Je größer die Korpora sind, desto mehr signifikante Zusammenhänge und Strukturen können in ihnen (bei vergleichbarer Streuung der Texte) aufgedeckt werden, desto genauere und detailliertere Untersuchungen können mit ihrer Hilfe durchgeführt werden und desto tragfähiger sind die Schlüsse, die aus den Auswertungen gezogen werden können.

Dies gilt insbesondere für die Untersuchung von seltenen Phänomenen, und das nicht nur, wenn seltene Wörter beteiligt sind, sondern vor allem auch, wenn eine Kombination von Wörtern und/oder Eigenschaften Gegenstand der Untersuchung ist. Möchte man z.B. den typischen Gebrauch eines Wor-

tes in seinem Kontext über eine längere Zeitreihe hinweg kontrastiv in verschiedenen Genres betrachten, so müsste ein zugrundeliegendes Archiv sehr umfangreich sein, um für jeden betrachteten Ausschnitt genügend Material zu bieten. So speziell das Beispiel scheinen mag, illustriert es doch sehr gut, dass es manchmal wichtiger ist, auf genügend Text mit ganz bestimmten Eigenschaften zurückgreifen zu können, als dass die Texte bezüglich der Eigenschaften eine schöne Verteilung im gesamten Korpus aufweisen. Oder anders ausgedrückt, nützt die beste Balanciertheit oder Ausgewogenheit nichts, wenn in der Schnittmenge der Kombination von textinternen und/oder -externen Eigenschaften keine oder nur wenige Texte verbleiben. Kenneth Church's Aussage (2003) „while balance is desirable, size is even more desirable" trifft ebenso zu, zumindest in der Hinsicht, dass bei der Erstellung eines Korpus heutzutage ein Zugewinn an Größe niemals zugunsten einer Vorstellung von Ausgewogenheit geopfert werden sollte – solange die technischen Rahmenbedingungen erfüllt sind (s. Kupietz i. Vorb.). Aufgrund dieser Überlegungen wird, wie oben ausgeführt, DeReKo kontinuierlich vor allem im Hinblick auf seine Größe, neben der Maximierung der Vielfalt durch die stratifizierte Stichprobenziehung, ausgebaut.

Metadaten und Annotationen

1991 wurde erstmals *ein* einheitliches Repräsentationsformat („BOT") für die Textstruktur und die Metadaten aller Korpustexte des IDS definiert und im Rahmen des Projekts COSMAS I für alle Korpustexte umgesetzt. Seine Grundeinheiten *Korpus, Dokument und Text* bilden auch heute noch das Gerüst des sog. IDS-Textmodells. BOT war textbasiert und beinhaltete neben einem Header-Bereich mit Metadaten auch Inline-Markup nach der sog. Mannheimer Konvention (MK). Das konkrete Markup für dieses Textmodell wurde im Laufe der Jahre einige Male an die Entwicklung der Annotationstandards angepasst. So wurde ab 1999 BOT/MK in den SGML-basierten Corpus Encoding Standard (CES, Ide 1998) überführt, 2006 dann in den XML-basierten Nachfolger XCES (Ide et al. 2000). CES basierte bereits auf dem TEI P3-Standard, und 2012/13 wurde eine neue, TEI P5-kompatible Dokumentgrammatik für das XCES-annotierte DeReKo spezifiziert („I5", vgl. Lüngen/Sperberg-McQueen 2012). Die Einhaltung standardisierter Annotationsformate erleichtert die Prüfung von Konsistenz und Integrität, garantiert die Interoperabilität von Ressourcen und ermöglicht beispielsweise die Teilnahme von DeReKo in Korpusverbünden wie CLARIN.

Die Metadaten in DeReKo umfassen nach dem IDS-Textmodell zunächst alle üblichen bibliografischen Angaben zur Text-, Dokument- und Korpusebene (Autor/Herausgeber, Titel, Untertitel, Verlag, Erscheinungsdatum und -ort) sowie als Besonderheit für die meisten Texte das Datum der Erstveröffentli-

chung und die (soweit bekannt) Entstehungszeit, die durchaus vom Erscheinungsdatum einer Publikation abweichen können (z.B. im Fall literarischer Werkausgaben), jedoch bei allen Untersuchungen von Sprachwandel über die Zeit eine wichtige Rolle spielen. Alle diese Metadaten können in COSMAS bei der Suche und bei der Zusammenstellung von virtuellen Teilkorpora genutzt werden. Des Weiteren bietet DeReKo Metadaten, die mittels Algorithmen und Skripts aus den Texten ermittelt werden, wie verschiedene deskriptiv-statistische Daten (Anzahl der Wörter, Tokens, Stoppwörter, Zahlen, Sätze, Absätze etc.), textlinguistische Kategorisierungen (Genre, Textsorte, Sachgebiet, Thema u.a.), geopolitische Zuordnung (D, AT, CH oder DDR), Angaben zur verwendeten Rechtschreibnorm sowie Angaben zu Duplikaten (vgl. Klosa et al. 2012).
Für die linguistische Annotierung wurde im Jahr 1993 der erste Schritt in eine neue Richtung gemacht, indem im Projektantrag des EU-Kooperationsprojekts MECOLB neben dem Handling von virtuellen Korpora und anderen innovativen Merkmalen auch die Verarbeitung von mehrfachen Annotationen in den Leistungsumfang des neu geplanten Korpusrecherche- und -analysesystems COSMAS II aufgenommen wurde. Vor dem Hintergrund, dass linguistische Annotationen immer eine bereits theorie- und implementierungsbedingte Interpretation der beobachteten Daten darstellen, galt das Ziel, DeReKo mit mehreren, möglichst verschiedenartigen Taggern auf verschiedenen linguistischen Ebenen mit ggf. sogar alternativen Annotationen pro Ebene zu versehen. Innerhalb von zwei Jahren gelang es dann, in Kooperation mit der Firma Logos alle damals verfügbaren Korpora mit einem von dem Logos-Übersetzungssystem abgeleiteten Tagger zu annotieren und im Jahr 1999 zusätzlich auch mit dem Gertwol Tagger der Firma Lingsoft Oy (Koskenniemi/Haapalainen 1996) mit weiteren linguistischen Annotationen anzureichern.
2007 wurde eine neue Tagging-Initiative gestartet. Ein Panel externer Experten wurde beauftragt, den gewachsenen Markt der Annotationstools zu sichten. Aus 25 sowohl kommerziellen als auch für den wissenschaftlichen Einsatz frei verfügbaren Werkzeugen und deren Beschreibungen wurden neun für eine tiefe Evaluation ausgewählt; aus der tiefen Evaluation nach einem Katalog linguistischer, technischer und ökonomischer Kriterien ging eine Empfehlung von drei Produkten hervor, für die in der Folge vom IDS Lizenzen erworben wurden, nämlich der TreeTagger von Helmut Schmid (Schmid 1994) und die beiden kommerziellen Systeme Machinese von Connexor Oy (Tapanainen/Järvinen 1997) und XFST Linguistic Suite von Xerox. In einer Studie (Belica et al. 2011) wurden die Annotationen der drei Werkzeuge im Hinblick auf ihre Übereinstimmung bzw. Abweichung evaluiert, und das Potenzial, aber auch mögliche Fallstricke der Verwendung morphosyntaktischer Annotationen bei der Korpusrecherche aufgezeigt. Aufgrund der Komplexität der Annotationen und ihrer XML-Repräsentation ist das annotierte

DeReKo heute 100-mal größer als die unannotierte Version. Die Wortarten-Annotationen des TreeTaggers und von Machinese können seit 2009 von den COSMAS-Nutzern bei der Korpusrecherche verwendet werden.

Lizenzmodelle und Musterverträge

Mitte 2004 wurden im Rahmen einer großen Akquisitionswelle, die auf Zeitungs- und Belletristik-Verlage abzielte, neue Muster-Lizenzvereinbarungen in Zusammenarbeit mit einer Anwaltskanzlei entwickelt. Basierend auf den bisherigen Erfahrungen, nach denen Lizenzverträge oft weder abgelehnt noch unterschrieben wurden, sollte vor allem die initiale Schwelle für die Unterzeichnung der Vereinbarungen durch die Lizenzgeber gesenkt werden und – angesichts der etwa 120 angeschriebenen Textgeber und der großen Anzahl der erhofften positiven Rückmeldungen – die Nachhaltigkeit der Vereinbarungen optimiert und der Aufwand für Folgevereinbarungen minimiert werden. Erreicht werden sollte dies u.a. durch eine jederzeitige teilweise und vollständige Kündbarkeit der Vereinbarung auch für bereits gelieferte Daten und einen generisch definierten Vertragsgegenstand im Sinne einer Rahmenlizenzvereinbarung, der es den Textgebern selbst überlies, ob diese überhaupt Daten gemäß der unterzeichneten Vereinbarung liefern würden, welche, wie viele und wie lange. Beide Schritte haben sich bewährt: Die Verweildauer der Verträge in den Lizenzabteilungen konnte deutlich reduziert werden, noch kein Lizenzgeber hat bisher von seinem Kündigungsrecht Gebrauch gemacht und fast alle der damals akquirierten Textgeber liefern heute noch Texte für DeReKo. Auch die

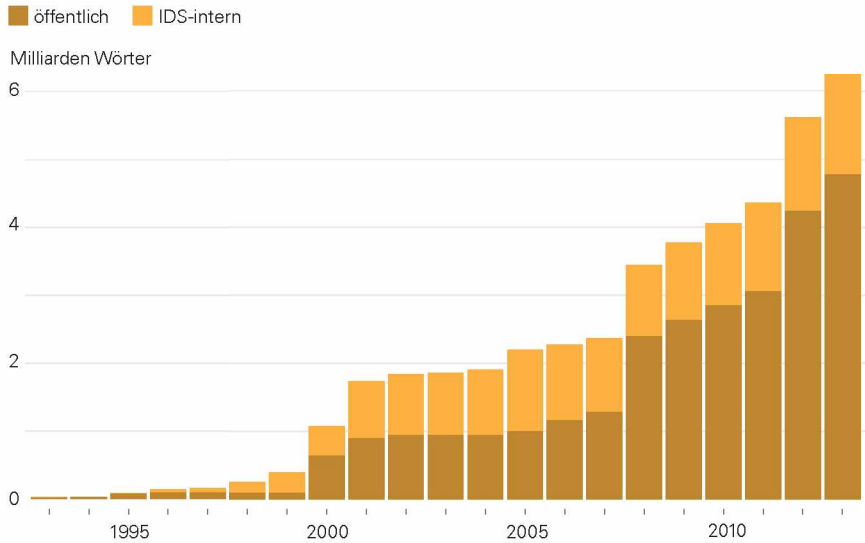

Abb. 2: Zunahme des DeReKo-Archivumfangs seit 1993

Lizenzvereinbarungen werden im Wesentlichen unverändert in ihren zwei Versionen, der von uns favorisierten, die auch die Nutzung der Texte von außerhalb des IDS erlaubt, und der Fallback-Version, die nur die Nutzung von Gästen und Mitarbeitern des IDS erlaubt, heute noch verwendet.
So konnte auch das wichtige Ziel, die Anzahl der zu verwaltenden Lizenzmodelle und den daraus resultierenden Konsequenzen für Software und Endnutzerlizenzvereinbarungen möglichst gering zu halten, erreicht werden. Die Lizenzverträge dienen heute außerdem auch als Musterlizenzverträge im CLARIN-Kontext.

Abb. 3: Die Textspender des Deutschen Referenzkorpus

Methodik der Korpusanalyse und -erschließung

Aus dem immensen Wachstum von Korpora im Allgemeinen und DeReKo im Besonderen ergeben sich neben neuen Möglichkeiten zuvor auch neue qualitative Herausforderungen. Die Größe der Korpora spielt einerseits eine entscheidende Rolle für die Untersuchung von sprachlichen Phänomenen in dünn besetzten sprachlichen Nischen. Andererseits ermöglicht bei sehr großen Belegmengen oft erst eine systematische Vorstrukturierung mit Hilfe von korpuslinguistischen Analysemethoden einen sinnvollen, d.h. Fehlinterpretationen vermeidenden Einblick in die Daten.

Die Erforschung und Entwicklung von korpuslinguistischen Analysemethoden und der Methodik zur Einbettung von empirischen Erkenntnissen in alle Bereiche der linguistischen Forschung bis hin zur Theoriebildung war bereits früh ein wichtiges Anliegen des IDS. Mit der Einrichtung des Projekts „Methoden der Korpusanalyse und -erschließung" im Jahr 2003 und des Programmbereichs „Korpuslinguistik" (2004) wurde der Raum für entsprechende Arbeiten institutionalisiert. Der Gegenstand des Projekts (und weitestgehend auch des Programmbereichs, s. Kupietz in diesem Band) ist zum einen die wissenschaftliche Erforschung der Methodik zur quantitativen und qualitativen Analyse von sehr großen Korpora, zum anderen die Modellierung der Prozesse linguistischer und allgemein kognitiver Interpretationen der in Sprache auftretenden Erwartbarkeiten. Die dabei gewonnenen Generalisierungen werden im Projekt auf wissenschaftstheoretischer Ebene reflektiert. Hierunter fallen neben neuen datengeleiteten Methoden auch hybride Ansätze, die verschiedene Datentypen miteinander kombinieren und sowohl datengeleitet als auch hypothesenbasiert oder in gleicher Weise mit Primärdaten und mit interpretativen Sekundärdaten, wie automatisch erzeugten linguistischen Annotationen, arbeiten (s. z.B. Müller 2007 und Kubczak/Konopka 2008).

Das zentrale Konzept zur Modellierung der präferenziellen Erwartbarkeiten stellen dabei Methoden der Kookkurrenzanalyse dar. Die Kookkurrenzanalyse eignet sich nämlich nicht nur als ein mächtiges Werkzeug zur Vorstrukturierung von großen Belegmengen bei interaktiver Korpusbefragung oder zur Extraktion vor festen Wortverbindungen, z.B. in der Phraseologieforschung, sondern sie eröffnet neue Möglichkeiten zur Gewinnung von latentem sprachlichen Wissen aus großen Korpora und bildet somit gleichzeitig auch den Ausgangspunkt eines innovativen Forschungsansatzes: Im Sinne des Paradigmas der „usage-based linguistics" (s. Bybee/Hopper 2001), nach dem linguistische Strukturen aus der Dynamik des Sprachgebrauchs hervortreten, wird das Systemische in der Sprache als emergentes Phänomen betrachtet. Mit Hilfe verschiedener auf Kookkurrenzanalyse basierender Ansätze zur Aufdeckung präferenz-relationaler Strukturen und ähnlichkeits-gruppieren-

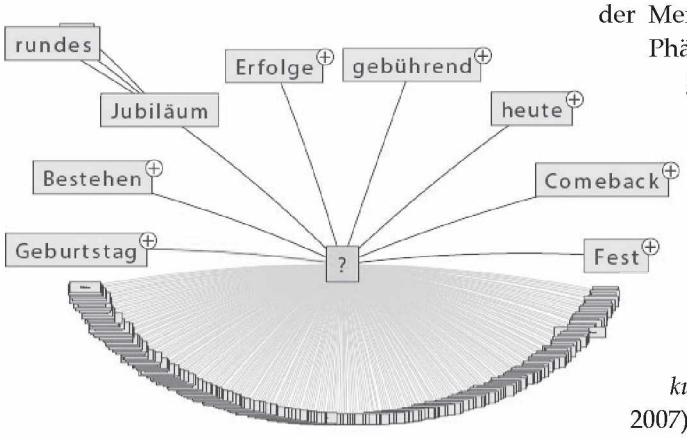

Abb. 4: (Berechnete) Kookkurrenzen lassen über ihre assoziative Wirkung sogar Bezugswörter (hier als Fragezeichen ausgeblendet) erraten

der Merkmale wird den Korrelaten dieser Phänomene im Sprachgebrauch nachgespürt.

Eine hochkomplexe Operationalisierung der Kookkurrenzanalyse ist z.B. seit 1995 ein fester Bestandteil der COSMAS-Plattform. Einen Einblick in ausgewählte Arbeiten aus diesem Bereich bietet als Denk- und Experimentierplattform die *Kookkurrenzdatenbank* CCDB (Belica 2001-2007). Sie bildet als große Sammlung von Kookkurrenzprofilen zu ca. 220.000 verschiedenen Lemmata eine solide Basis, um die im Sprachgebrauch manifesten emergenten Strukturen systematisch aufzudecken, zu inventarisieren, zu interpretieren und theoretisch zu begründen. Die Sammlung enthält zu jedem Lemma die Ergebnisse von Kookkurrenzanalysen in Form von Hierarchien von ähnlichen Verwendungen, sog. Kookkurrenzprofile, mit bis zu 100.000 Verwendungsbeispielen pro Lemma und Analyse. Über den Vergleich von Kookkurrenzprofilen lassen sich Verwandtschaftsrelationen zwischen den Lemmata erschließen, in Verbindung mit Clusteringverfahren ermöglicht diese die Kartierung des Verwendungsspektrums einzelner oder von Paaren von Lemmata zur Deutung interner oder bilateraler Bedeutungsaspekte.

Ein Desiderat ist in diesem Bereich noch ein umfassendes generisches Vorgehensmodell, das den empirisch arbeitenden Linguisten bei der Ausgestaltung seines Untersuchungsszenarios leitet: Von der Korpuskomposition und der Wahl der Methoden über die Verknüpfung der Erkenntnisse mit dem empirischen Material bis hin zu deren Dokumentation ggf. integriert in ein zu publizierendes Werk. Teilaspekte eines derartigen Modells sind prototypisch für DeReKo-Daten und die Kookkurrenzanalyse beschrieben und operationalisiert (Perkuhn 2007a, 2007b). Die von der Analysemethode vorgeschlagenen Strukturen werden für verschiedene Explorationstechniken visualisiert angeboten. Gewonnene Erkenntnisse können ergänzend zu diesen Strukturen dokumentiert werden und stehen für weitere Interpretations- und Verwertungsschritte zur Verfügung.

Während das datengetriebene Analyseparadigma bis vor einiger Zeit vor allem für die Lexikologie relevant war, können heute auf der Grundlage sehr großer Stichproben wesentlich komplexere sprachliche Muster und Strukturen aufgedeckt und auch in Abhängigkeit anderer Faktoren (z.B. Zeit, Herkunft)

analysiert werden (s. Keibel et al. 2008). Dies zeigt sich zurzeit nicht nur in aktuellen Tendenzen in der Grammatikforschung, die z.B. in der neuen Konferenzreihe *Grammar and Corpora* (Štícha/Fried 2008 und Konopka et al. 2011) vorgestellt werden, sondern auch in der linguistischen Theoriebildung insgesamt, etwa durch neue Zeitschriften wie *Corpus Linguistics and Linguistic Theory*.

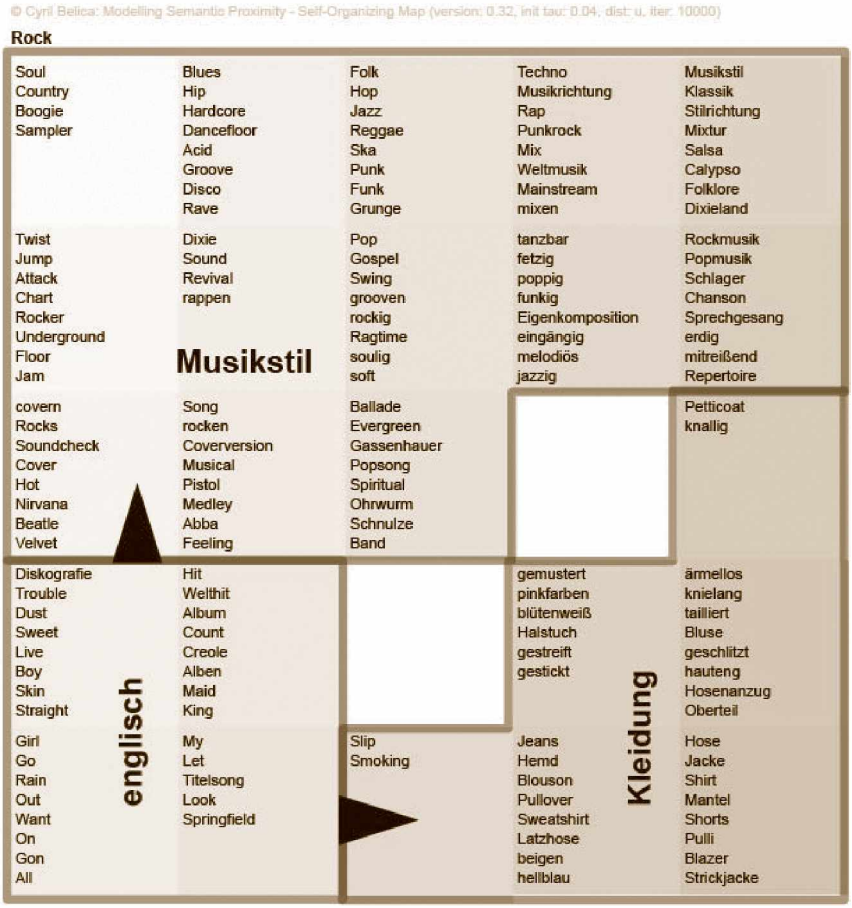

Abb. 5: Bedeutungsaspekte des Wortes „Rock", visualisiert anhand (erst automatisch, dann händisch) gruppierter verwendungsähnlicher Wörter

Dank

Wir danken den derzeitigen Mitarbeitern des Programmbereichs Korpuslinguistik am IDS Marc Kupietz, Harald Lüngen und insbesondere Rainer Perkuhn für ihre konstruktiven Anregungen zum zweiten Teil dieses Beitrags. Rainer Perkuhn danken wir auch für die redaktionelle Zusammenführung der beiden Teile.

Literatur

- **Bankhardt, Christina** (2009): D-Spin – Eine Infrastruktur für deutsche Sprachressourcen. Sprachreport 1/2009, S. 30-31.
- **Belica, Cyril** (2001-2007): Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. Mannheim. Internet: http://corpora.ids-mannheim.de/ccdb/.
- **Belica, Cyril et al.** (2011): The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls. In: Konopka, Marek et al. (Hg.), S. 451-469.
- **Biber, Douglas** (2008): Representativeness in Corpus Design. In: Fontenelle, T. (Hg.): Practical Lexicography. Oxford, S. 63-88.
- **Bybee, Joan L./Hopper, Paul J.** (Hg.) (2001): Frequency and the Emergence of Linguistic Structure. (= Typological Studies in Language 45). Amsterdam.
- **Church, Kenneth W.** (2003): Speech and language processing: Where have we been and where are we going? In: Proceedings of the 8th European conference on speech communication and technology, 1-4. Genf. Internet: www.isca-speech.org/archive/eurospeech_2003/e03_0001.html.
- **Church, Kenneth W./Mercer, Robert L.** (1993): Introduction to the special issue on computational linguistics using large corpora. Computational Linguistics 19 (1), S. 1-24.
- **Gietz, Peter et al.** (2006): Textgrid and ehumanities. In: Proceedings of the Second IEEE International Conference on e-Science and Grid Computing E-SCIENCE '06, Amsterdam. IEEE Computer Society.
- **Ide, Nancy** (1998): Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. Proceedings of the First International Language Resources and Evaluation Conference, Granada, S. 463-470.
- **Ide, Nancy,/Bonhomme, P./Romary, L.** (2000): XCES: An XML-based encoding standard for linguistic corpora. In: Proceedings of the Second International Language Resources and Evaluation Conference (LREC'00). Paris.
- **IDS** (Hg.) (1967): Satz und Wort im heutigen Deutsch. Probleme und Ergebnisse neuerer Forschung. Jahrbuch 1965/1966. Düsseldorf.
- **IDS** (2013): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2013-I (Release vom 19.3.2013). Mannheim. Internet: www.ids-mannheim.de/kl/projekte/korpora/archiv.html.
- **Keibel, Holger/Kupietz, Marc/Belica, Cyril** (2008): Approaching grammar: Inferring operational constituents of language use from large corpora. In: Štícha, František/Fried, Mirjam (Hg.), S. 235-242.
- **Klosa, Annette/Kupietz, Marc/Lüngen, Harald** (2012): Zum Nutzen von Korpusauszeichnungen für die Lexikographie. In: Lexicographica 28, S. 71-97.
- **Konopka, Marek et al.** (Hg.) (2011): Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.-24.9.2009. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 1). Tübingen.

- **Koskenniemi, Kimmo/Haapalainen, Mariikka** (1996): GERTWOL – Lingsoft Oy. In: Hausser, Roland (Hg.): Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994. (= Sprache und Information 34). Tübingen, S. 121-140.
- **Kubczak, Jacqueline/Konopka, Marek** (2008): Grammatical Variation in Near-Standard German: a corpus-based project at the Institute for the German Language (IDS) in Mannheim. In: Štícha, František/Fried, Mirjam (Hg.), S. 251-260.
- **Kupietz, Marc et al.** (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, N. et al. (Hg.): Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), S. 1848-1854. Internet: www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.
- **Kupietz, Marc** (i. Vorb.): Constructing a Corpus. In: Durkin, Philip (Hg.): The Oxford Handbook of Lexicography. Oxford.
- **Lüngen, Harald/Sperberg-McQueen, Michael** (2012): 'A TEI P5 Document Grammar for the IDS Text Model. In: Journal of the Text Encoding Initiative (2012). Internet: http://jtei.revues.org/508.
- **Müller, Stefan** (2007): Qualitative Korpusanalyse für die Grammatiktheorie: Introspektion vs. Korpus. In: Kallmeyer, Werner/Zifonun, Gisela (Hg.): Sprachkorpora – Datenmengen und Erkenntnisfortschritt. Berlin/New York, S. 70-90.
- **Perkuhn, Rainer** (2007a): Systematic Exploration of Collocation Profiles. In: Proceedings of the 4th Corpus Linguistics Conference (CL 2007), Birmingham. Internet: www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/132Paper.pdf.
- **Perkuhn, Rainer** (2007b): „Corpus-driven": Systematische Auswertung automatisch ermittelter sprachlicher Muster. In: Kämper, Heidrun/Eichinger, Ludwig M. (Hg.): Sprach-Perspektiven. Germanistische Linguistik und das Institut für Deutsche Sprache. (= Studien zur Deutschen Sprache 40). Tübingen, S. 465-491.
- **Schmid, Helmut** (1994): Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing. Manchester, UK, S. 44-49.
- **Štícha, František/Fried, Mirjam Fried** (Hg.) (2008): Selected contributions from the conference Grammar and Corpora, Sept. 25-27, 2007, Liblice, Czech Republic. Prag.
- **Tapanainen, Pasi and Timo Järvinen** (1997): A non-projective dependency parser. In: Proceedings of the 5th Conference on Applied Natural Language Processing. Washington DC, S. 64-71.
- **van Uytvanck, Dieter** (2010): CLARIN Short Guide on Virtual Collections. Technical report, CLARIN. Internet: www.clarin.eu/files/virtual_collections-CLARIN-ShortGuide.pdf.
- **Váradi, Tamás et al.** (2008): CLARIN: Common language resources and technology infrastructure. In: Proceedings of the 6th International Language Resources and Evaluation Conference (LREC'08), Paris. European Language Resources Association (ELRA).